OD BIOGRAFSKEGA LEKSIKONA DO ZNANSTVENOKRITIČNE IZDAJE: VPRAŠANJE TRAJNOSTI ELEKTRONSKIH BESEDIL

Tomaž Erjavec
Jan Jona Javoršek
Matija Ogrin
Petra Vide Ogrin

Oddano: 27. 10. 2010 – Sprejeto: 23. 2. 2011

Pregledni znanstveni članek
UDK 004.91:801.7(497.4)

Izvleček

Članek predstavlja probleme digitalnega kuratorstva vključno z izbiro formata dokumenta, sistema za predstavitev in avtorskih pravic, in sicer v okviru dveh primerov projektov digitalizacije obstoječih pisnih virov: *Elektronske znanstvenokritične izdaje slovenskega slovstva* ter *Slovenskega biografskega leksikona*. Posveča se trem osnovnih vidikom digitalnega kuratorstva, ki se dotikajo ohranitve virov. Prvi vidik je trajnost digitalnega formata, torej zapisa besedila; na tem področju smo se oprli na mednarodne standarde in sprejete prakse, zlasti smernice iniciative TEI (*Guidelines for Electronic Text Encoding and Interchange*), ki je samodokumentativen, razviden in široko sprejet standard. Drugi vidik je predstavitev gradiva in iskalnik, kjer smo uporabili statični HTML, digitalno knjižnico Fedora Commons in iskalnik SoLR. Zadnji vidik pa je vprašanje avtorskih pravic in dostopa do gradiva: v kakšnem formatu, komu in pod kakšnimi pogoji je gradivo dostopno. Za pričujoče digitalne izdaje je bil kot najbolj merodajni argument izbran vidik kar največje odprtosti, kakor jo določa licenca Creative Commons.

Ključne besede: digitalno kuratorstvo, digitalne izdaje, digitalne knjižnice, TEI, XSLT, Fedora Commons

ERJAVEC, Tomaž; Jan Jona JAVORŠEK; Matija OGRIN; Petra VIDE OGRIN. From biographical lexicon to scholarly edition: the question of sustainability of digital editions. Knjižnica, 55(2011)1, pp. 103–114

Review article
UDC 004.91:801.7(497.4)

Abstract

The paper deals with the issues of digital curation, including storage formats, presentation, and access rights, in the frame of the case-study of two projects on digitisation of written materials: the *Scholarly digital editions of Slovenian literature* and the *Slovenian biographical lexicon.* Three basic aspects of digital curation, including preservation, are discussed. The first aspect is the sustainability of our digital format, i.e. of our text-encoding; where the international standards and best practices are used, particularly the TEI *Guidelines for Electronic Text Encoding and Interchange* as a self-documenting, transparent and widely adopted de-facto standard. The second aspect is the presentation and search over the materials, where the static HTML pages and the Fedora Commons repository with SoLR full-text search are used. The last aspect deals with the access rights to the materials: in what format and for whom, and under what conditions are the texts made available. The argument on maximum openness as is embodied in the Creative Commons a licence was adopted as most applicative solution for the presented digital editions.

Keywords: digital curation, digital edition, digital libraries, TEI, XSLT, Fedora Commons

1 Uvod

V sodelovanju med Znanstvenoraziskovalnim centrom Slovenske akademije znanosti in umetnosti (ZRC SAZU), Institutom »Jožef Stefan« (IJS) in Slovensko akademijo znanosti in umetnosti (SAZU) je v zadnjem desetletju nastala vrsta elektronskih znanstvenih izdaj in besedilnih korpusov. Te izdaje so vidna plat obsežnega raziskovalnega dela s področja več humanističnih ved, zlasti literarne in obče zgodovine ter jezikoslovja. Leksikoni ter znanstvenokritične izdaje predstavljajo dela s kar največjo gostoto podatkov in s kar najbolj kompleksno členitvijo besedila. Razumljivo je, da so bila zato za ohranitev in posredovanje obsežnega in dolgotrajnega raziskovalnega dela, ki je omogočilo nastanek teh publikacij, pretehtano izbrana tudi najboljša sredstva, metode in protokoli za njihovo trajnost in diseminacijo.

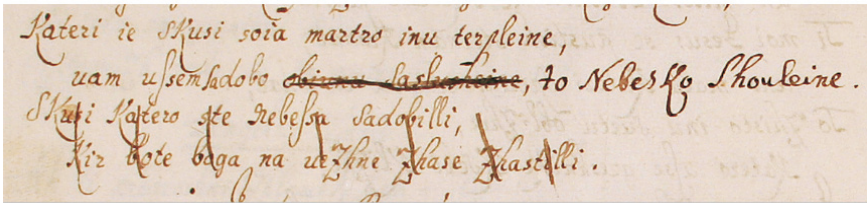
Ker je problematika trajnosti digitalnih besedil večplastna (Kenney in McGovern, 2003), se bomo v tem prispevku na študiji dveh konkretnih primerov osredotočili predvsem na zagotavljanje standardov zapisa in preverjanje njegove pravilnosti, kar digitalni zapis besedil varuje pred zastaranjem. V drugem delu besedila bomo predstavili še tehnološko podporo za prikaz in iskanje po podatkih, zaključili pa bomo s problematiko dostopa do integralnih besedil, kar vključuje vprašanje avtorskih pravic nad digitalnimi izdajami.

2 Zvrsti in oblike besedil

Projekta, katerih rezultati so že dostopni uporabnikom in ki implementirata zgoraj omenjene vidike trajnosti, sta naslednja: Elektronske znanstvenokritične izdaje slovenskega slovstva (eZISS) in Slovenski biografski leksikon 1925–1991 (SBL).

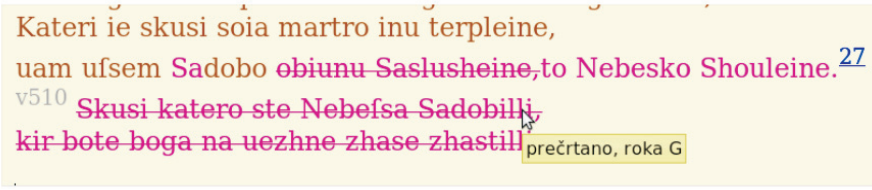
Zametki projekta **Elektronske znanstvenokritične izdaje slovenskega slovstva – eZISS** (http://nl.ijs.si/e-zrc/) segajo v leto 2001. Njegova osnovna ambicija je bila raziskati, kako v elektronski medij prenesti ter aplicirati klasične metode, ki veljajo za področje priprave znanstvenokritičnih izdaj. Treba je bilo preučiti, katere tehnologije omogočajo podrobno znanstveno strukturiranje besedil, primeren znanstveni aparat, vzporedne prikaze več metodološko različnih prepisov besedila, vzporeden prikaz faksimila in prepisa, vključitev zvočnih datotek za reprodukcijo fonetičnega prepisa itn. Ta vprašanja so bila uspešno razrešena tako teoretično kakor praktično z dosledno uporabo zapisa XML po smernicah konzorcija Text Encoding Initiative (TEI, http://www.tei-c.org) (Guidelines, 2007) – glede na realne potrebe in posebnosti posamezne izdaje. Prva med njimi – *Tri pridige o jeziku* A. M. Slomška – je izšla v letu 2004. Odtlej je do leta 2010 skupno izšlo sedem izdaj, okvirno po ena na leto, in obsegajo od srednjeveških in baročnih rokopisov do modernega romana (Iz. Cankar, *S poti*) in poezije (Podbevšek). Izdaja z najbolj kompleksno notranjo strukturo so *Brižinski spomeniki,* ki obsega poleg faksimilov, slovarja, študij idr. glavno besedilo spomenikov v kar 17 različnih prepisih in prevodih, ki jih je moč pregledovati v vzporednem prikazu, po želji s posebno tipografijo ZRCola ali po standardnem Unikodu. S tem je projekt eZISS izkoristil potencial Smernic TEI za nadaljevanje evropske filološke tradicije v digitalnem kontekstu (Burnard, 2005) – tudi za najbolj kompleksne in pretanjene besedilne strukture, ki jih morejo postaviti zahteve kritičnih izdaj.

Kako je zahtevnejša besedilna struktura zajeta in eksplicirana v znanstvenokritični izdaji, je prikazano na spodnjih primerih: Slika 1 prikazuje fragment *Škofjeloškega pasijona*, Slika 2 prikazuje isto besedilo v elektronski znanstvenokritični izdaji, kakor ga vidimo na zaslonu, medtem ko Slika 3 prikazuje podstat elektronske izdaje, ki prikaz na zaslonu sploh omogoča: strukturirano besedilo, podrobno označeno z elementi XML TEI, kjer so vsebinske informacije (denimo



Slika 1: Fragment Škofjeloškega pasijona

paleografski podatki o rokah, ki so pisale v rokopis), eksplicitno izražene na standardiziran način.

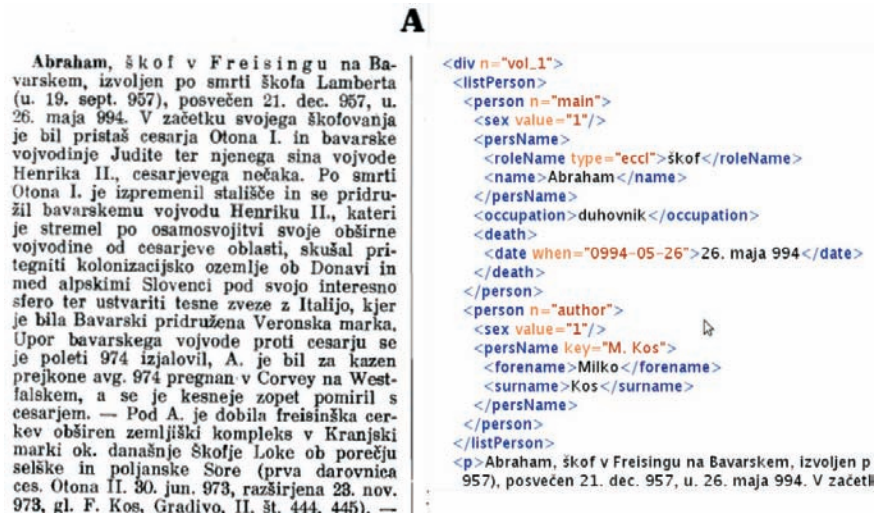


Slika 2: Zaslonski prikaz fragmenta *Škofjeloškega pasijona* v elektronski znanstvenokritični izdaji

```
3
4  <l n="508" xml:id="sp_dt.l.508">Kateri ie skusi soia martro inu terpleine,</l>
5  <l n="509" xml:id="sp_dt.l.509">uam ufsem <add hand="#hB" place="inline">Sa</add>dobo
6    <subst status="unremarkable">
7      <del hand="#hG">obiunu Saslusheine,</del>
8      <add hand="#hG" place="inline">to Nebesko Shouleine.</add>
9    </subst><note xml:id="sp_dt.note.92" resp="#MOG" place="end">Poseg neznane roke [...] </note></l>
10 <l n="510" xml:id="sp_dt.l.510">
11   <del hand="#hG">Skusi katero ste Nebefsa Sadobilli,</del></l>
12 <l n="511" xml:id="sp_dt.l.511">
13   <del hand="#hG">kir bote boga na uezhne zhase zhastilli.</del> </l>
```

Slika 3: Fragment *Škofjeloškega pasijona* v strukturiranem zapisu XML TEI

Slovenski biografski leksikon (1925–1991, SBL) je osrednji in najpomembnejši znanstveni rezultat slovenske nacionalne leksikografije. Nastajal je skoraj sedem desetletij, vanj so bile vključene vsaj tri generacije vodilnih slovenskih humanistov, mnogi njegovi članki pa so ne le pregled védenja o določeni osebi, ampak več – so temeljne raziskave, ki so utrle novo območje znanja. SBL je po 2. svetovni vojni izhajal pri SAZU in ZRC SAZU. Ker so bili njegovi prvi, predvojni zvezki že bibliofilska redkost, se je Biblioteka SAZU odločila, da SBL ponovno izda, in sicer kot elektronsko publikacijo. Samo besedilo SBL je sicer ostalo nespremenjeno, vendar pa smo t. i. metapodatke o predstavljenih osebnostih izluščili iz člankov in jih podrobno strukturirali – prav tako po smernicah TEI. S tem se je vsako biografsko geslo spremenilo v dvoje: v besedilo izvornega članka in v blok strukturiranih metapodatkov o osebi (Slika 4). Te metapodatke, ki so namenjeni strojnemu iskanju, smo izdatno elaborirali, usklajevali in dopolnjevali. Njihova podrobna strukturiranost z oznakami TEI omogoča, da je prek iskalnika po njih mogoče opravljati zahtevnejše poizvedbe z raznimi iskalnimi pogoji.



A

Abraham, škof v Freisingu na Bavarskem, izvoljen po smrti škofa Lamberta (u. 19. sept. 957), posvečen 21. dec. 957, u. 26. maja 994. V začetku svojega škofovanja je bil pristaš cesarja Otona I. in bavarske vojvodinje Judite ter njenega sina vojvode Henrika II., cesarjevega nečaka. Po smrti Otona I. je izpremenil stališče in se pridružil bavarskemu vojvodu Henriku II., kateri je stremel po osamosvojitvi svoje obširne vojvodine od cesarjeve oblasti, skušal pritegniti kolonizacijsko ozemlje ob Donavi in med alpskimi Slovenci pod svojo interesno sfero ter ustvariti tesne zveze z Italijo, kjer je bila Bavarski pridružena Veronska marka. Upor bavarskega vojvode proti cesarju se je poleti 974 izjalovil, A. je bil za kazen prejkone avg. 974 pregnan v Corvey na Westfalskem, a se je kesneje zopet pomiril s cesarjem. — Pod A. je dobila freisinška cerkev obširen zemljiški kompleks v Kranjski marki ok. današnje Škofje Loke ob porečju selške in poljanske Sore (prva darovnica ces. Otona II. 30. jun. 973, razširjena 23. nov. 973, gl. F. Kos, Gradivo, II. št. 444, 445). —

```
<div n="vol_1">
 <listPerson>
  <person n="main">
   <sex value="1"/>
   <persName>
    <roleName type="eccl">škof</roleName>
    <name>Abraham</name>
   </persName>
   <occupation>duhovnik</occupation>
   <death>
    <date when="0994-05-26">26. maja 994</date>
   </death>
  </person>
  <person n="author">
   <sex value="1"/>
   <persName key="M. Kos">
    <forename>Milko</forename>
    <surname>Kos</surname>
   </persName>
  </person>
 </listPerson>
 <p>Abraham, škof v Freisingu na Bavarskem, izvoljen p
  957), posvečen 21. dec. 957, u. 26. maja 994. V začetk
```

Slika 4: Primer biografskega članka v tiskanem SBL in njegove pretvorbe v strukturiran zapis XML TEI, kjer so podatki iz implicitne preneseni v eksplicitno strukturo

Kakor je razbrati že iz teh bežnih opomb, že zgoraj omenjena projekta obsegata nabor na moč raznolikih besedil. Raznolike so tako njihove zvrsti kakor viri, ki jih predstavljajo. Struktura znanstvene izdaje baročnega rokopisnega teksta, kakršen je *Škofjeloški pasijon*, ki je povrh kot literana zvrst drama s svojimi posebnimi elementi od didaskalij do govorcev, se zelo razlikuje od strukture izdaje modernega romana *S poti*. Vse izdaje eZISS pa imajo povsem drugačen notranji sestav kakor *Slovenski biografski leksikon*. Za vse te strukture – od rokopisnih vrinkov poznejše roke med vrsticami rokopisa do biografskih podatkov, kakor so razne vrste in sestavine osebnih imen, razne vrste datumov in krajevnih imen, poklici itn. – je bilo treba najti ustrezno označevanje in ga konsistentno uporabljati, izdaje pa tudi predstaviti na spletu in na ustrezen način ponuditi v uporabo.

3 Uporaba standardov

Za trajnost teh elektronskih besedil je, kar zadeva standarde elektronskega označevanja oz. zapisa, poskrbljeno s tem, da so vse omenjene besedilne edicije označene v zapisu XML po smernicah TEI (Guidelines, 2007). Smernice TEI, s polnim imenom (ang.) *The TEI Guidelines for Electronic Text Encoding and Interchange*, definirajo in dokumentirajo označevalni jezik za predstavitev strukturnih, pred-

stavitvenih ali konceptualnih lastnosti besedil. Priporočila so usmerjena predvsem v zapis dokumentov s področja humanistike in družboslovja, še najbolj pa v zapis primarnih virov za namene raziskav in analiz. Izražena so kot modularna in razširljiva shema XML, ki ji je pridružena podrobna dokumentacija, dostopna pa so pod odprtokodno licenco. Smernice TEI so šle skozi več izdaj; zadnja je TEI P5 iz leta 2007, ki je usklajena z ustreznimi smernicami W3C in ISO, ki jih *ipso facto* upoštevajo tudi naše izdaje.

Pri označevanju dokumentov smo tako uporabljali elemente in atribute XML, ki jih določa TEI, pri čemer je pomembno, da imajo ti elementi in atributi pripisano semantiko, dokumentirano v smernicah, ki jo je potrebno dosledno upoštevati pri označevanju. V nekaj (redkih) primerih, kjer smernice niso vsebovale ustrezne rešitve za opis, pa smo uporabili razširitveni mehanizem, ki ga ravno tako predpisujejo smernice, in obstoječo shemo razširili s svojimi elementi.

Kanonična oblika naših besedil je tako dokument XML, ki je veljaven (validiran) glede na shemo XML, ki je narejena v skladu s smernicami TEI. S tem je zagotovljeno, da neodvisno in ločeno od naših datotek obstaja podrobna dokumentacija – smernice TEI –, v kateri je struktura naših besedil dokumentirana tako s proznim opisom (smernice) kakor s shemo XML (Relax NG), s pomočjo katere je mogoče pravilnost strukture teh besedil kadar koli preveriti.

Uporaba TEI npr. omogoča formaliziran način predstavitve časa življenja (in dela) ljudi iz SBL; ta, na prvi pogled preprost podatek (datum rojstva in smrti), se izkaže kot problematičen v primerih, kakor so »v prvi polovici XV. stoletja«, »ne po 1650«, pri čemer ima TEI atribute, ISO pa format za izražanje takšnih časov in trajanj. Poleg samih smernic ponuja TEI tudi razna orodja za delo z besedili, predvsem transformacije XSLT za pretvorbo besedil TEI v HTML, ki so opisane v naslednjem poglavju.

4 Predstavitev in iskanje

Izdaje eZISS so na splet postavljene na razmeroma enostaven način: izvorni TEI XML vsake izdaje je s transformacijami XSLT (ki je priporočilo W3C) preveden v datoteke HTML, in te so statično postavljene na splet. Za prvih nekaj izdaj smo napisali svoje transformacije XSLT, za vsako izdajo posebej, za novejša pa smo privzeli standardne pretvorbe TEI XSLT.

Ta pristop ne omogoča poglobljenega iskanja po besedilih, je pa primeren za branje. Vsaka izdaja je torej digitalna knjiga, ki ima pripisane naslove URI (ang. uniform resource identifier); tako naj bi – v idealnem primeru za vedno – deseto

poglavje vzporednega prikaza romana *S poti* Izidorja Cankarja iz knjižnice eZISS imelo sledeči URL http://nl.ijs.si/e-zrc/izidor/html/spoti-APP.html#body.div.11, URL odstavka 226 (prav tam) pa bi se končal s #p.226. Tako je zagotovljen stalen naslov URL za vir, ki je vedno dostopen na istem mestu, ta naslov pa ostane vedno veljaven tudi za navajanje. Ta preprosta lastnost, ki jo pričakujemo od vseh spletno objavljenih dokumentov, je žal le redko izpolnjena (Berners-Lee, 1992), vendar je še posebej pomembna za digitalne objave. Privzeti način pa ima tudi slabost, da dostopa do starih različic istega besedila ne podpira neposredno, saj ima vsako besedilo ne glede na različico pripisan en sam URI.

Pristop s statično transformacijo XML v HTML je razmeroma starokopiten, saj vnaprej postavi vizualizacije posameznih elementov TEI in medsebojne povezave med strukturami. Ne omogoča torej prilagodljivih parametrov prikaza, dinamičnega povezovanja vsebin in strukturiranega iskanja po besedilih. Če je to za izdaje eZISS še sprejemljivo, pa je dosti manj za referenčne priročnike, kakršen je SBL.

Zato smo se odločili za ambicioznejšo spletno izdajo, in sicer na osnovi platforme Fedora Commons (http://www.fedora-commons.org/). Ta platforma, ki se je v zadnjih letih razvila kot platforma za digitalne knjižnice, se je zdela ustrezna izbira predvsem zato, ker ima veliko razvojno bazo, je ustrezno vzdrževana in v stalnem razvoju, ima dobro pripravljeno in dostopno dokumentacijo, je prosto dostopna tudi v izvorni obliki in razvita na temelju sprejetih standardov in praks (Lagoze et al., 2006).

Osnovni gradnik digitalne knjižnice na platformi Fedora Commons je dokument, ki ga s stališča implementacije razumemo kot objekt v pomenu objektno-orientiranega programiranja. Vsak dokument je zapisan kot dokument v zapisu XML, ki poleg sistemskih komponent, kakršne so identifikator (ID) dokumenta ipd., vsebuje vrsto podatkovnih tokov (ang. datastreams) oz. elementov dokumenta, ki so vsi vključeni dokumenti XML z lastnimi imenskimi prostori, in vrsto diseminatorjev (disseminators), ki so opisani v jeziku za opis spletnih servisov WSDL (in v kontekstu objektno-orientiranega programiranja delujejo kot objektne metode). Dejanska implementacija je dokaj elegantna, saj sistemski podatkovni tokovi vključujejo metapodatke Dublin Core, odnose med dokumenti (izraženi so kot trojčki v zapisu RDF – (ang.) resource description framework) ter popis sprememb dokumenta (ang. audit trail), tako da so sistemski podatki zapisani in dostopni na enak način kakor vsebina dokumentov. Poleg tega imajo podatkovni tokovi tudi različice, kar pomeni, da ob popravljanju že objavljenega dokumenta ne pobrišemo prejšnje različice, ki ostane dostopna in se je mogoče še vedno sklicevati nanjo. Tudi diseminatorji so zapisani na podoben način, vendar so v modelu digitalnih objektov platforme zapisani v posebnih dokumentih, na katere se dejanski dokumenti sklicujejo z odnosom (kar omogoča implementacijo objektno-orientiranega razreda objektov, kjer imajo vsi člani razreda na voljo

vse metode, ki jih razred implementira). Tako ob pripravi zbirke dokumentov zadošča, da diseminatorje določimo enkrat, vsi drugi dokumenti v zbirki pa se nanje sklicujejo. Dejanska implementacija je nekaj kompleksnejša, ker loči abstraktno definicijo servisov (objektno-orientiranega vmesnika razreda) od opisa implementacije, kar omogoča izmenjavo objektov med različnimi namestitvami digitalne knjižnice z različnimi implementacijami servisov. Poleg tega model omogoča tudi sklicevanje na zunanje podatkovne tokove z naslovom URI, pri čemer je zunanji vir lahko drug dokument v digitalni knjižnici ali spletna stran na poljubnem strežniku.

Platforma vsebuje tudi vrsto servisov, med katerimi so najpomembnejši filter za transformacije XSLT (XML stylesheet transform), iskalnik po metapodatkih (Dublin Core), iskalnik po odnosih RDF (ki sprejema različne standardne jezike za povpraševanja in omogoča uporabo vrste standardnih formatov za odgovore) ter vmesnik za vključevanje zunanjih iskalnih sistemov in protokol za zajem metapodatkov iniciative Open Archive (OAI-MH).

Platforma je kljub relativni kompleksnosti konceptualno enostavna in zelo fleksibilna. Objekti so v veliki meri samozadostni, uporaba diseminatorjev in filtrov XSLT pa omogoča različne načine prikaza podatkov. Še več, ker je mogoče naslov posameznega diseminatorja uporabiti kot naslov URI zunanjega podatkovnega toka, imajo dokumenti lahko tudi dinamične podatkovne tokove, poleg tega pa lahko diseminatorji, ki kot vhodne podatke vedno uporabljajo podatkovne tokove dokumenta, na ta način uporabijo rezultate drugega diseminatorja. Tako je mogoče graditi verige transformacij XSLT oz. uporabljati platformo kot sistem za upravljanje s podatkovnimi tokovi XML.

Pri implementaciji digitalne knjižnice smo si postavili nekaj preprostih ciljev:

a. Vsi podatki in prikazi na spletnih straneh digitalne knjižnice morajo biti transformacije ustrezno oblikovanih dokumentov TEI, ki morajo biti dostopni prek spletne strani.
b. Čim več označenih podatkov mora biti na spletu ustrezno prikazanih.
c. Označeni podatki in odnosi med dokumenti morajo biti aktivni, kar pomeni, da moramo omogočiti iskanje, urejanje ipd. na podlagi teh podatkov.
d. Za potrebe metapodatkovne izmenjave je treba čim več metapodatkov zajeti v Dublin Core, čeprav ostajajo merodajni podatki v zapisu TEI (zlasti <teiHeader>), ki omogočajo bogatejši in točnejši zapis.
e. Podatki morajo ostati dostopni tudi mimo sistema, tako da je mogoče po potrebi preiti na drugo platformo, in sicer tudi v primeru, če bi aplikacija odpovedala. (Slednje Fedora Commons omogoča preprosto zato, ker so vsi podatki v dobro dokumentiranih datotekah XML dostopni v datotečnem sistemu, različne baze podatkov pa služijo samo kot orodje za hitrejši dostop do teh istih podatkov.)

f. Podatkom je treba prirediti trajni identifikator, na osnovi katerega je mogoče določiti trajni naslov vira (URL).

Da bi lahko dosegli vse cilje, smo platformo prek vmesnika za iskalne sisteme razširili z iskalnikom Lucene SoLR. Tako smo dobili zelo močan iskalni sistem, ki uporablja transformacije XSLT za zajem podatkov in omogoča iskanje s poizvedovalnim jezikom knjižnice Lucene, pri čemer SoLR omogoča tudi bogat prikaz iskalnih rezultatov, ki ga je mogoče s transformacijami XSLT tudi prilagoditi uporabniškemu vmesniku.

Pri organizaciji digitalne knjižnice smo se držali vodila, da en dokument TEI ni nujno tudi en sam dokument digitalne knjižnice, temveč je treba uporabiti najbolj primerno osnovno enoto glede na vsebino ter s pomočjo odnosov med dokumenti v knjižnici ustrezno izraziti dejansko organizacijo podatkov. V nekaterih primerih, npr. pri digitalnih objavah strokovnih ali leposlovnih besedil, ostanejo ta razmerja enostavna, saj eno besedilo, digitalizirano kot en dokument TEI, ostane en dokument tudi v digitalni knjižnici, pa čeprav ga vmesnik morda prikazuje deljeno, npr. tako da uvodni material (kolofon, uvod, spremno besedo, uredniške opombe) in kritiški aparat prikažemo z dodatnim klikom na povezavo, potem ko se je na prvi strani dokumenta prikazal začetek dejanskega besedila. Na enak način je mogoče v sistemu prikazovati tudi obstoječe izdaje, npr. izdaje eZISS, preprosto tako, da navedemo naslov izdaje v obliki TEI kot zunanji podatkovni tok dokumenta.

Zato pa je v nekaterih drugih primerih, med njimi je trenutno edina javno dostopna izdaja *Slovenski biografski leksikon,* to razmerje nekoliko bolj zapleteno. Pri SBL smo za enoto, ki je v digitalni zbirki postala posamezen dokument, izbrali članek biografskega leksikona. Vsi članki imajo odnos »sem element zbirke«, ki kaže na osnovni dokument, ta pa vsebuje uvodni in zaključni aparat ter elemente, ki so potrebni za poizvedovalni sistem. Poleg tega imajo posamezni članki tudi odnose, ki kažejo na naslednji in prejšnji članek (za listanje), knjigo izvirne izdaje (za določanje datuma izida ipd.) ter črko, ki ji pripadajo (za abecedne sezname).

Že to zadošča za ustrezno spletno predstavitev, ki omogoča udobno brskanje tudi po tako zapleteni publikaciji, kakršen je referenčni priročnik. Vendar seveda to ni dovolj, saj je zaradi narave publikacije in bogastva podatkov, ki so bili digitalizirani in ustrezno označeni, nujno omogočiti aktivno uporabo podatkov s povezovanjem in iskanjem. To nalogo je opravil iskalnik Lucene SoLR. Iskalnik zgradi indekse ključnih besed, po katerih je mogoče iskati, in sicer na podlagi filtra XSLT, ki dokument preoblikuje v seznam ključnih besed oziroma zaporedij ključnih besed. V okviru platforme Fedora Commons je to najudobneje doseči tako, da vsakemu dokumentu pripišemo filter XSLT, ki zajame iskalne ključe iz ustrezno označenih elementov (v primeru *Slovenskega biografskega leksikona* so to zlasti elementi znotraj elementa <Person>, ki opisujejo podatke o posamezni

osebi), pri čemer je mogoče določeno ključno besedo, npr. »besedilo«, uporabiti tudi za zajem celotnega besedila člankov, in že imamo na voljo poizvedovanje po vsem besedilu.

Pri tem je najpomembnejše, da lahko pri poizvedovanju uporabljamo posamezne ključne besede v okviru poizvedovalnega jezika Lucene, kar pomeni, da lahko leksikon poprosimo za seznam vseh oseb, denimo, ki so moškega spola, rojene 1850–1900 v Ljubljani in so delovale na določenem področju (kajti vse dejavnosti so urejene v hierarhičen sistem oz. taksonomijo prav zaradi tovrstnih poizvedovanj).

Iskalni sistem se je izkazal kot izjemno fleksibilen in uporaben, tako da je v spletni publikaciji uporabljen na tri načine: dve različni maski omogočata dva načina iskanja (preprosto iskanje, kjer v okence napišemo besedo ali izraz v jeziku Lucene, ter si s klikom ogledamo zadetke), ter kompleksnejši način, kjer si pomagamo z večjo iskalno masko, ki za nas zgradi ustrezen iskalni izraz), poleg tega pa je iskalni sistem uporabljen tudi za vrsto praktičnih povezav, npr. »Rojeni / Umrli na današnji dan« na uvodni strani publikacije ali »Sodobniki« in »Generacija« na straneh posameznih člankov. Sistem uporabljamo tudi za druge priročne sezname, npr. za prikaz vseh člankov posameznega avtorja leksikona ipd.

Odzivi testnih uporabnikov so bili dobri, tako da razvijamo novo elektronsko izdajo *Slovenskega biografskega leksikona,* ki bo v iskalnem sistemu omogočila uporabo še več označenih podatkov v najnovejši digitalizirani izdaji v obliki TEI, ter pod imenom *Slovenska biografija* povezala obstoječo publikacijo z digitalizacijo *Primorskega slovenskega biografskega leksikona* ter novim *Slovenskim biografskim leksikonom 2* (v pripravi). Poleg tega tudi širimo digitalno knjižnico, saj bo v kratkem pod naslovom *Elektronska znanstvena besedila* poleg *Slovenske biografije* vključevala tudi eZISS, eZMono (*Elektronske znanstvene monografije*) ter vrsto specializiranih izdaj različnih raziskovalnih projektov. Vse te publikacije bodo imele lastne vmesnike in iskalne sisteme v okviru iste platforme Fedora Commons, poizvedovati po njih pa bo mogoče tudi skozi vmesnik celotne digitalne knjižnice. Na ta način upamo, da bomo dosegli tako strokovnemu kakor poljudnemu uporabniku čim bolj prijazno obliko digitalne izdaje, ne da bi se oddaljili od strogih meril strokovne digitalne forme.

Za trajnost spletnih izdaj je seveda pomembna tudi stabilnost strežnika, na katerem so izdaje dostopne; vse naše izdaje domujejo na strežniku nl.ijs.si, ki praktično nepretrgoma obratuje že petnajst let; sistem je osnovan na odprti kodi, torej Linux, Apache HTTPD, Apache Tomcat itn., in je lociran v sodobnem, ustrezno opremljenem podatkovnem centru.

5 Avtorske pravice

Za maksimalno odprtost naših izdaj za večino dokumentov uporabljamo licence Creative Commons (http://creativecommons.org/); te licence omogočajo bistveno več svobode pri kopiranju, nadaljnjem razširjanju in predelavi del kakor pa klasična formulacija avtorskih pravic, kjer so vse te pravice uporabniku odvzete. Tako lahko zainteresirani izdaje eZISS ne samo berejo v HTML prek spleta, pač pa lahko celotne izdaje, vključno z izvornimi datotekami XML in transormacijami XSLT, uporabljenimi za pretvorbo v HTML, uporabniki brez izpolnjevanja pogodb kopirajo na svoj računalnik in v večini primerov tudi predelujejo glede na svoje potrebe, prav tako pa omogočajo tretjim osebam dostop do (predelanih) izdaj.

S tem pristopom želimo kar najbolj povečati uporabnost svojega dela, posredno pa skrbimo tudi za ohranjanje teh izdaj, saj bo delo hranjeno in dostopno v več kopijah oz. različicah. Hkrati velja omeniti, da je težnja po trajnosti in obči dostopnosti naših elektronskih besedil notranje povezana z njihovo pomembnostjo in kvaliteto, saj se trudimo za objavo temeljnih, referenčnih del kulture slovenskega naroda, ki imajo trajno vrednost.

6 Zaključki

Prispevek je predstavil dosedanje elektronske izdaje, ki so nastale kot plod sodelovanja ZRC SAZU in IJS. Osredotočili smo se na uporabo standardov pri zapisu materialov, njihovo postavitev na splet, orodja za prikaz in iskanje ter avtorske pravice nad digitalnimi besedili. Za trajnost digitalnih izdaj je poskrbljeno z več vidikov: pred zastaranjem zapisa nas varuje dosledna uporaba široko uveljavljenjih odprtih standardov zapisa (Unikod, XML, TEI), pred fizičnim propadom pa na eni strani postavitev materialov na dobro vzdrževan strežnik v okviru IJS, na drugi pa prosto razširjanje izvorne oblike digitalnega gradiva.

Glede zadnjega vidika – nevarnosti fizičnega uničenja datotek – bi bilo dobrodošlo, če bi v svojih projektih lahko zagotovili dodatno kopiranje digitalnega gradiva v oddaljen repozitorij za dolgoročno hrambo, kakršne imajo nekatere nacionalne knjižnice. Te rešitve trenutno nimamo, kar se zdi šibka točka. Vendar smo želeli poudariti, da je v vsakdanji praksi za trajnost elektronskih besedil pomembnejši prvi vidik: standardizirano kodiranje besedil – saj v praksi več besedil propade zaradi nestandardnih oblik zapisa kakor zaradi fizičnega uničenja datotek. Na tem področju sta projekta izdaj eZISS in SBL naredila bistvene korake tako glede metod izdelave kakor glede trajnosti samega zapisa.

Navedeni viri

1. Berners-Lee, T. (1992). Cool URI's don't change. V T. Berners-Lee, *Style guide for online hypertext*. Cambridge: W3C Consortium. Pridobljeno 10. 9. 2010 s spletne strani: http://www.w3.org/Provider/Style/URI
2. Burnard, L. (2005). Encoding standards for the electronic edition. V M. Ogrin (Ur.), *Znanstvene izdaje in elektronski medij* (str. 25–42). Ljubljana: ZRC SAZU.
3. *Guidelines for electronic text encoding and interchange.* (2007). Charlottesville: TEI Consortium. Pridobljeno 10. 9. 2010 s spletne strani: http://www.tei-c.org/Guidelines/P5/
4. Kenney, A. R. in McGovern, N. Y. (2003). The five organizational stages of digital preservation. V P. Hodges (Ur.), D*igital libraries: a vision for the 21 century: a festschrift in honor Wendy Lougee* (str. 122–153). Ann Arbor: The Scholarly Publishing Office, University of Michigan.
5. Lagoze, C., Payette, S., Shin, E. in Wilper, C. (2006). Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries*, *6* (2), 124–138.

Dr. Tomaž Erjavec je zaposlen na Institutu »Jožef Stefan«.
Naslov: Jamova cesta 39, 1000 Ljubljana
Naslov elektronske pošte: tomaz.erjavec@ijs.si

Jan Jona Javoršek je zaposlen na Institutu »Jožef Stefan«.
Naslov: Jamova cesta 39, 1000 Ljubljana
Naslov elektronske pošte: jona.javorsek@ijs.si

Dr. Matija Ogrin je zaposlen v Znanstvenoraziskovalnem centru SAZU.
Naslov: Novi trg 2, 1000 Ljubljana
Naslov elektronske pošte: matija.ogrin@zrc-sazu.si

Mag. Petra Vide Ogrin je zaposlena na Slovenski akademiji znanosti in umetnosti.
Naslov: Novi trg 3, 1000 Ljubljana
Naslov elektronske pošte: petra.vide@zrc-sazu.si